

What Is the Relevance of Sequencing Primate Genomes to Understand Human Disease?

Ethan Wan

Received May 22, 2024

Accepted August 14, 2024

Electronic access September 15, 2024

Background: Human diseases are caused by a combination of genetic, environmental and lifestyle factors. Clinical understanding of genetic variants and their impact on diseases is crucial in the development of precision medicine. However, identifying pathogenic genetic variants is challenging and inconclusive using only human genomes databases. This paper aims to examine and answer the questions: whether and why the NHP (non human primates) genome sequencing can help understand human disease, and how to use the data effectively.

Methods: The paper reviews multiple research documents found from respected science research journals and government health agency websites to consolidate understanding and synthesize the information in a logical interconnected order with key findings and conclusions.

Results: The research findings include: identifying pathogenic human genetic variants is essential to advance individualized precision medicine, where NHP genome is valuable to help understanding such genetic variants relevance; and how projects such as PrimateAI-3D trained AI machine learning models with NHP genome data and helped improve clinical interpretation of human genetic variants.

Conclusions: This research paper reviews how the sequencing of NHP genomes is being used with AI technologies as a valuable genome dataset to advance human health research. For future medical research, it is worthwhile to scale more comprehensive sequencing of NHP genomes and optimize AI machine learning models for more effective processing.

Keywords: NHP(non-human primates), genomes, precision medicine, PrimateAI-3D, AI, machine learning

Introduction

Background and Context: With the advancement in genomes sequencing, molecular biology, computer science and medical research, clinical understanding of a genetic variant's impact on diseases becomes crucial to help predict, prevent and treat diseases more effectively with individualized approaches.

Problem Statement and Rationale: However, with current enormous human genome databases, actionable accurate clinical interpretation of genetic variants remains challenging to achieve. Researchers recently began to use NHP genome data to help analyze, annotate and assess human genetic variants. This research aims to find out whether and why NHP genomes data is a helpful data source to use and how.

Significance and Purpose: This research confirms NHP genome data is a valuable data set to use with AI processing to help achieve more effective interpretation of human genetic variants. The research also advocates for more comprehensive NHP genome sequencing and more AI models to be built to optimize NHP genome processing and generate further insights into human genetics.

Objectives: The research reviews and synthesizes relevant

science literature aiming to answer the guiding questions: whether NHP genome data is useful in helping understanding human disease, and if yes, why and how? What are effective methods to use this data and what do these mean for future research direction to achieve better understanding of human genetic variants?

Scope and Limitations: The study reviewed multiple research documents and publications mostly from recent 2-3 years in respectable science journals and sources including Science, Nature, NEJM, NIH. The documents covered topics regarding interconnection of human genetic data and precision medicine and target drug development, and how NHP genome data helps in clinical interpretation of human genetic data, including details of the PrimateAI-3D project based on the PrimeAI-3D team's publications in Science journal. One limitation is the performance benchmark of PrimateAI-3D is sourced from a complete yet unpublished journal from medRxiv, authored by the PrimateAI-3D project team instead of by an objective 3rd party, so it might be biased and need further peer review and 3rd party benchmark testing.

Methodology Overview: The research is done by searching in respectable science journals (such as Science, Nature, NEJM),

NCBI databases and NIH and other government research agency websites using keywords related to NHP genomes data, human genetic variant interpretation, precision medicine and target drug development. The information is synthesized and consolidated to reach the conclusions answering original research questions.

Results

Genetic Variants Interpretation in Human Medical Research

DNA (Deoxyribonucleic Acid), the core template of genetics, is a molecule with an intricate structure encoding the genetic instructions which regulate the development and function of all living organisms. For each species, the DNA encoding's unique genetic signature is characterized by a specific sequence and number of nucleotides, which scientists have named the genome¹.

To elucidate the genetic code of a living organism, and to uncover the entire list of nucleotides making up its genome, scientists have derived complex laboratory techniques called DNA sequencing to capture, purify, and read DNA molecules one nucleotide at a time. Through the sequencing process, researchers can accurately identify each nucleotide and its position in a given DNA molecule, thus reconstructing the genomic sequence of the organism. Global cooperation paved the path for the success of the Human Genome Project, which eventually sequenced for the first time the entire human genetic code. The completion of the first draft of the human genome in 2003 was the most significant breakthrough that revolutionized human biology and genetic studies².

Since then, scientists have embarked on a new era in molecular biology and medical research. New tools and methodologies in molecular biology help analyze mechanisms of living organisms; sequencing of entire human genomes makes it possible to identify correlation of genetic variants with phenotypes, AI and computer science advancement help processing enormous genome data which was impossible with human manual labor. The convergence of advancements in all these areas contribute to the development of personalized and individualized medicine, also called precision medicine, the new medical paradigm and approach based on each individual's genetic profile, and use the understanding and interpretation of genetic variants to help predict, prevent and treat diseases more proactively³.

As part of natural evolution, DNA can have mutations, or variants. A mutation can be silent and evolutionary neutral, benign and harmless, or harmful and potentially causing diseases. For example, a missense mutation, also called a nonsynonymous mutation, changes an amino acid in a protein. A nonsense mutation changes an amino acid in a protein to a stop codon which ends synthesis of the protein at that location. The most harmful mutation is frameshift mutation, where insertion

and deletion of nucleotides can change the genetic code that tells the cell which amino acid to put into the protein at that location, hence causing loss of protein function. With millions of human individuals having their genomes sequenced, the enormous different genetic variants shown among individuals are hard to interpret and their pathogenic risk are hard to assess, as most of individual genetic variants are relatively new in evolutionary history and haven't gone through long natural selection, there is no sufficient history and data to help identify whether an individual genetic variant is benign or harmful with potential to cause diseases.

Identifying pathogenic genetic variants can help in multiple stages in precision medicine based on an individual's genetic profile: 1). Predict if an individual has genes that are prone to causing certain diseases; 2). Prevent chronic disease proactively by using preventive medicine or therapy and adopting more healthy lifestyle changes; 3). Helping the discovery, development and usage of a new target medicine, e.g. helping material design of a drug to target a specific gene, helping clinical trial optimization of a new target drug, and treating a patient based on the specific genetic profile to make individualized decisions on which target drug or therapy to use and what dosage.

Precision medicine is turning the old "one size fits all" approach to an individualized approach in diagnostics and drug therapy and prevention. Human beings are similar, but also individually different. A medicine applied to different individuals might have different impacts. Genomics shows at a molecular level differences between individuals and helps make individual predictions about disease risk that can help somebody choose an appropriate prevention plan. It also allows the possibility of picking the right drug at the right dose for the right person³.

Traditionally, medical practice has been reactive and is focused on curative treatment and practiced on the scale of a general population. The new approach takes account of an individual's specific characteristics, such as genetic and epigenetic profile, to help assess genetic and environmental contribution to their health, and predict risk and prevent diseases when a human is still healthy and not suffering from a disease yet, by giving treatment and care to prevent a potential illness, especially for chronic diseases where a patient does not die immediately but suffer a negative impact on their quality of life. The new approach aims to not only treat the symptoms, but also the underlying causes of the disease, with greater emphasis on disease prevention and promoting overall health, and select the most appropriate treatments based on the individual's genetic, phenotypic and lifestyle characteristics.

Drug targets refer to the key molecules involved in certain metabolic pathways that result in diseases. For example, in the oncology field, targeted drugs can block or turn off signals that make cancer cells grow, or can signal the cancer cells to self destroy. Researchers are developing more targeted

drugs while learning more about specific changes in cancer cells. Understanding the underlying genetics can facilitate the successful development of new therapies.

So in summary, Interpreting the impact of genetic variants on human health is essential in personalized genomic medicine, especially in genetic risk prediction and drug target discovery.

In the past decade, GWAS (genome-wide association studies) have identified tens of thousands of common variants associated with phenotypes and diseases. Each associated variant usually can only explain a very small percentage of a phenotype trait. Multiple associated variants can be combined into a PRS (polygenic risk score) to explain a considerable portion of disease risk. However, it is challenging to identify specific genes as pathogenic from GWAS because most GWAS variants reside in the noncoding part of the genome. In contrast, rare variant studies connect a specific variant directly to clinical phenotypes. However, while for rare genetic disorders and cancers, rare variants are routinely examined to explain why a disease occurred, study of rare variants for common diseases have not progressed much because of imprecise interpretation of variant function and insufficient cohort sizes for rare variant analysis⁴.

Many human genomes databases are created and publicly accessible for science and medical research, such as gnomAD, ClinVar and UKBiobank. The gnomAD (Genome Aggregation Database) is developed by an international coalition aiming to aggregate and harmonize both exome and genome sequencing data from many large-scale sequencing projects, and making data available for the wider scientific community. ClinVar is another freely accessible public archive of human genetic variations classified for diseases and drug responses. ClinVar data contains clinically asserted relationships between human variation and observed conditions, and the assertion history⁵.

In addition, machine learning and AI (Artificial Intelligence) techniques as well as advanced statistical techniques are playing an important role in helping analyze all these genomes data. However, with currently available massive data of human genomes, comprehensive understanding of human genetic variants and accurate identification of pathogenic variants, is still challenging.

NHP Genomics Research

In order to decipher the human genome, search for genetic variants which might cause human diseases or disease susceptibility and develop targeted personalized treatment mechanisms, sifting through human genomes alone can be an uphill battle to detect disease-related gene variants. To better understand and process the human genome datasets, researchers began to compare the human genome to the DNA of other living organisms. Comparative genomics is a research field that aims to examine the commonalities and differences

between genomes of different species. With the comparison, researchers can gain a more detailed view of the complex genetic factors underlying molecular, cellular, and biological processes. Particularly, sequencing of primate genomes, human's closest biological relatives, can help gather valuable additional datasets and perspectives.

Primates, including humans, gorillas and chimpanzees, are a diverse order of mammals with unique anatomical and physiological traits and features which differentiate them from other mammalian species. First, primate brain size is relatively larger in proportion to their body weight, generally believed to contribute to more complex cognitive capabilities and social network behaviors absent in most animals. Second, primates have opposable thumbs on their hands and feet and flat nails instead of claws, enabling adroit object handling, an essential anatomical feature for primate-specialized activities like grooming and grabbing. Further, the brain region dedicated to vision, the occipital lobe, is more developed in primates than any other mammal, and the migration of their eye orbits to the anterior part of the face allows them to have a binocular vision. As NHPs are humans' closest relatives, studying NHP biology and especially their genetics provides a massive amount of information which could facilitate a better understanding of human biology and diseases⁵.

After *Homo sapiens*, the chimpanzee became the first primate species whose genome was sequenced in 2005, and after that, many other primate species with various degrees of relatedness to humans have had their genome sequenced as well.

Insights gathered by comparing NHP genomes to human genomes are leveraged to facilitate the discovery of the genetic basis of human diseases and provide insights into the molecular mechanism underlying human health and wellbeing. Through comparative genomics, researchers compare genetic history and adaptation, similarities and differences between humans and primate genomes, also specific genetic mutations associated with various diseases and assess their prevalence in humans. Finding genetic variables affecting both resilience and susceptibility to a disease is made easier by researching the genetic diversity of primate species⁶.

There are two major reasons that NHP genome data is useful as complementary datasets to use in understanding human genetic variants. As a result of the short evolutionary distance between humans and nonhuman primates, human proteins share near-perfect amino acid sequence identity with NHPs. Hence, the effects of a protein-changing mutation found in one primate species are likely to be concordant in the other primate species⁷.

Another important reason is NHP genomes have gone through tens of millions of years of natural selection, while human DNA records the evolutionary history of a few hundred thousand years. By systematically cataloging common variants of NHPs, scientists aimed to annotate these variants as being unlikely to cause human disease as they are tolerated by natural selection

in a biologically closely related species. The annotation of such benign common genetic variants can be super valuable to be used to infer the effects of variants across the genome using machine learning. Bringing in NHP genomes as complementary data sets in training AI and machine learning models is tremendously helpful to improve clinical understanding of human genetic mutations⁷.

PrimateAI-3D: NHP Genomes Sequencing, Machine Learning Algorithm and Results

PrimateAI-3D is a project created by the biotech company Illumina, to analyze NHP genomes via AI and machine learning to help identify benign and pathologic variants in human genetics. The PrimateAI-3D team sequenced massive NHP genomes data, then purify, transform and converge the NHP genomes as source dataset to train machine learning models in their AI convolutional neural network, which then processes multiple human genomes databases to help identify pathogenic variants.

In collaboration with the international science community, scientists at Illumina recently published the results of their study of primate genomes in four papers in the Science journal. The study sequenced over 800 individuals from 233 species of NHPs, representing all 16 families and over 86% of living genera, then developed and used PrimateAI-3D to process the sequenced data. PrimateAI-3D is an algorithm built on deep machine learning language and architecture similar to those used in ChatGPT, but designed to model genomic sequences. The strategy and supporting hypothesis behind PrimateAI-3D is that with the long history of natural selection, many common genetic variants in NHPs can be ruled out for causing disease, and be annotated as benign and used effectively to train the AI models parameters. The AI neural network learns where benign variants are represented in a gene and, with process of elimination, which regions are likely to be pathogenic if mutation occurs. In this way, it learned how to accurately predict pathogenic human genetic variants better than only using human genome data since there is no clear clinical understanding of most genetic variants in the latter.

The diagram below is cited directly from “The landscape of tolerated genetic variation in humans and primates” published in June 2023 in Science journal, illustrating the deep learning models that PrimateAI-3D trained using common NHP genetic variants identified as benign⁷:

Following the strategy outlined above, researchers obtained whole-genome sequencing for 809 individuals from 233 NHP species and cataloged 4.3 million common missense variants. Researchers confirmed that human missense variants seen in at least one NHP species were annotated as benign in the ClinVar database in 99% of cases. In contrast, common variants from mammals and vertebrates outside the primate lineage

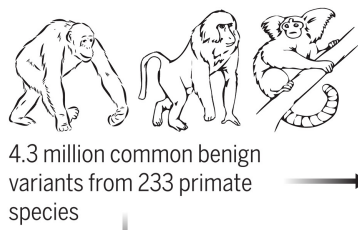
were substantially less likely to be benign in the ClinVar database (71 to 87% benign), restricting this strategy to NHPs. Overall, researchers reclassified more than 4 million human missense variants of previously unknown impact as likely benign, increasing the number of annotated missense variants to greater than 50-fold compared to existing clinical databases⁷. In addition, by not relying on annotations from existing clinical variant databases, and assigning benign annotation to common primates variants, PrimateAI-3D manages to provide an unbiased review of variant pathogenicity.

To identify the pathogenicity of the remaining missense variants in the human genome, researchers in Illumina constructed PrimateAI-3D as a semi-supervised machine learning system that uses a convolutional neural network (CNN) operating on voxelized protein structures. Below is the PrimateAI-3D architecture diagram directly cited from the Science paper which Illumina published in 2023⁷:

CNN is an artificial neural network and a system of hardware and software patterned simulating the way how human brains learn and process images. As one of the most popular and powerful types of deep learning algorithms. CNNs are specifically designed to map image data to output variables and by processing training sets of annotated images, the machine can also learn to identify elements which are characteristic of objects within the images⁸. Unlike deep-learning architectures that relied on linear protein sequence, PrimateAI-3D uses 3D convolutions to recognize key structural and evolutionary patterns in proteins, the network is also taught to predict masked amino acids from the surrounding 3D context, a technique borrowed from language models that are trained to predict missing words in sentences. Researchers trained PrimateAI-3D to separate common primate variants from matched control variants in 3D space as a semi-supervised learning task. In another separate task, language models and multiple sequence alignments are also trained to incorporate evolutionary amino acid constraints across diverse species.

The PrimateAI-3D team evaluated the trained PrimateAI-3D model alongside 15 other published machine learning methods on the ability to distinguish between benign and pathogenic variants in six different clinical benchmarks and claimed that PrimateAI-3D outperformed all other classifiers⁷. PrimateAI-3D compares 233 primate genomes to half a million individual human genome samples and identifies genomic similarities between NHPs and humans, which allows it to determine nucleotide sequences prone to be pathogenic⁷.

Their earlier work used 385,000 missense variants from 134 individuals across six primate species, then scaled up to 4.5 million primate missense variants by sequencing another 703 individuals across 211 NHP species. The selected species represent approximately half of the 521 extant primate species and cover all major primate families. They targeted an average of 3.5 individuals per species to ensure that common variants

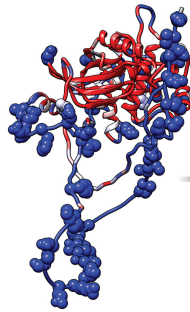


4.3 million common benign variants from 233 primate species

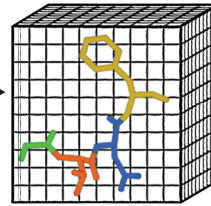
Validation of primate variants in human clinical variant database



98.7% of common primate variants in ClinVar are benign

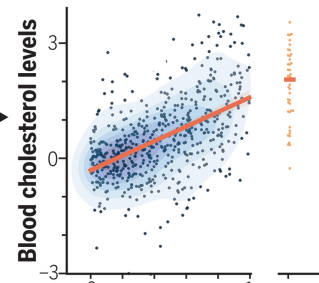


Benign primate variants superimposed on 3D protein structures

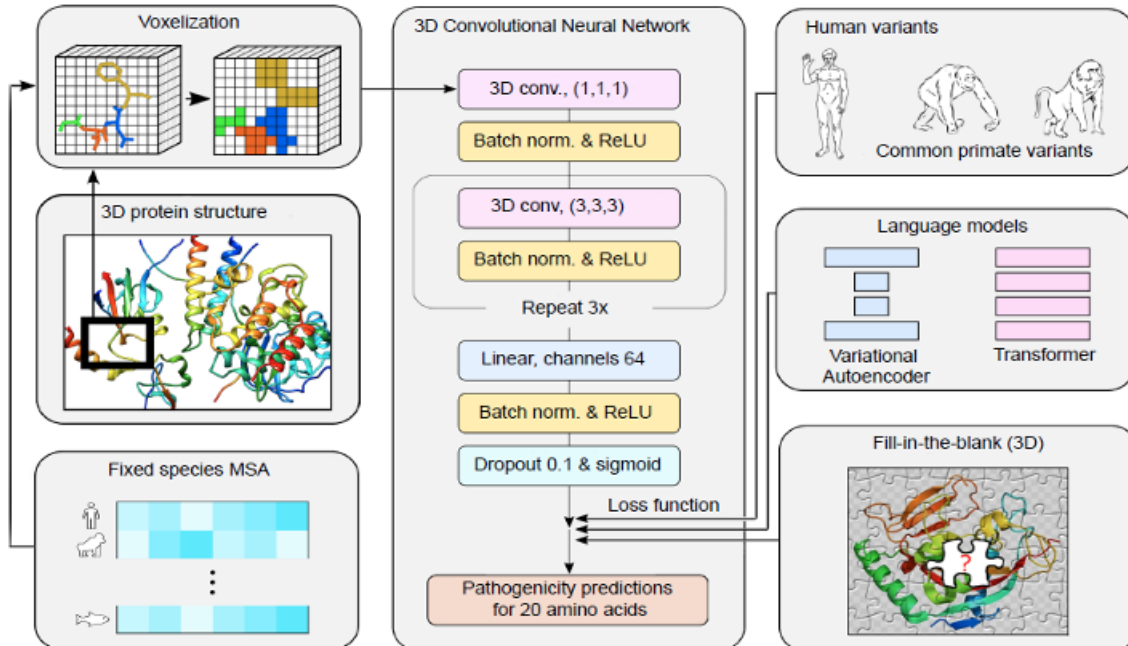


3D convolutions + deep learning protein language models

Individuals carrying LDLR variants



PrimateAI-3D score **LoF**
Validation of variant effect predictions in clinical cohorts



are sampled rather than rare mutations. As a result of the scaled sequencing, they tremendously increased the number of common variants used to train a machine-learning classifier⁹.

In an Illumina article “Improving genetic risk prediction and drug target discovery using primate DNA and advanced artificial intelligence”⁹, it explains how PrimateAI-3D uses machine learning and AI to help drug target research. This state-of-the-art classifier accurately quantified missense variant pathogenicity in humans, and improved discovery of genes impacting clinical phenotypes. PrimateAI-3D integrated rare and common variant PRS models into a unified risk score to provide a more comprehensive understanding of disease risk, and risk scores in rare variant analysis successfully identified genes as candidate drug targets, including known examples such as PCSK9, HMGCR (target of lipid-lowering statins), ANGPTL3, and NPC1L1⁹.

Cholesterol-related diseases are a significant concern due to their impact on cardiovascular health. Two key proteins involved in cholesterol metabolism are PCSK9 (Proprotein Convertase Subtilisin/Kexin Type 9) and HMGCR (3-hydroxy-3-methylglutaryl-CoA reductase). PCSK9 is a protein that regulates the number of “bad cholesterol” receptors on cells, which in turn influences the amount of “bad cholesterol” in the bloodstream. Inhibiting PCSK9 can lead to a reduction in LDL cholesterol levels, potentially lowering the risk of cardiovascular diseases. HMGCR is an enzyme that plays a crucial role in the cholesterol synthesis pathway. Statins drugs, commonly used to mitigate high cholesterol, work by inhibiting HMGCR, thereby reducing cholesterol production in the liver⁹. Both PCSK9 and HMGCR inhibitors have been instrumental in the management and prevention of cholesterol-related diseases, highlighting the importance of understanding and targeting these proteins in cardiovascular health. PrimateAI-3D validated the diagnostic power of its algorithm, by identifying the linkage of genes such as PCSK9 and HMGCR to cholesterol-related diseases. Besides PCSK9 and HMGCR, PrimateAI-3D also identified other proteins as candidate drug targets, e.g. ANGPTL3 as a drug target in hyperlipidemia and atherosclerosis and NPC1L1 as another drug target relevant in cholesterol treatment. These findings highlight the potential of using PrimateAI-3D scores in rare variant analyses to uncover new drug targets⁹.

The study also compared PrimateAI-3D against 15 published missense pathogenicity prediction methods and claimed PrimateAI-3D outperformed all other classifiers by accurately distinguishing pathogenic variants from benign across four cohorts—the UK Biobank, a neurodevelopmental disorders cohort (DDD), an autism spectrum disorders cohort (ASD), and a congenital heart disease cohort (CHD). In addition, PrimateAI-3D was the best classifier for separating benign and pathogenic variants in the ClinVar database and had the highest average correlation with deep mutational scan assays¹⁰.

The following figure demonstrating the PrimateAI-3D

performance is cited from the article “PrimateAI-3D outperforms AlphaMissense in real-world cohorts” published by David Parry et al from Illumina AI lab¹⁰:

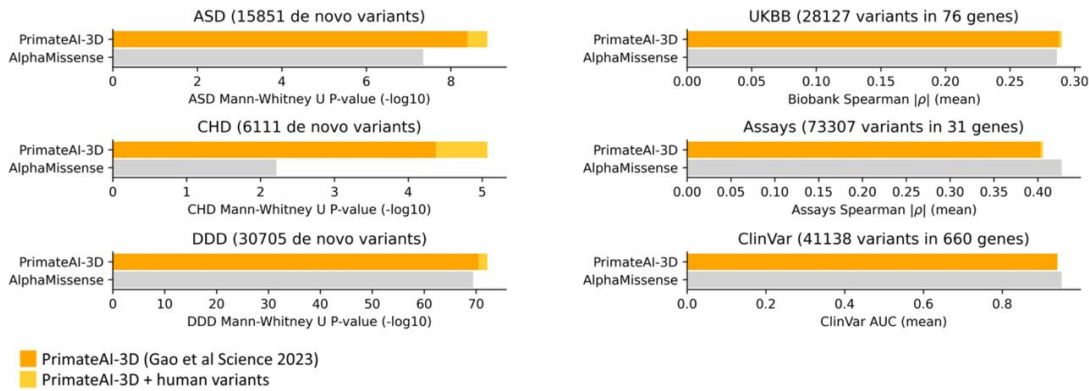
Illumina researchers prepared the most comprehensive set of benchmarks to date and applied them to the recently published models from PrimateAI-3D and AlphaMissense and claimed PrimateAI-3D outperformed AlphaMissense in all four large clinical cohorts consisting of half a million individuals from DDD, ASD, CHD and UKBioBank¹⁰.

AlphaMissense was created by researchers at DeepMind, one of Google’s AI companies, to predict how dangerous a genetic mutation is to help research into rare diseases¹¹. AlphaMissense used clinically annotated variants from the ClinVar database for model selection and hyperparameter training. AlphaMissense shares many similarities with PrimateAI-3D, including training the model using common primate variants as benign versus mutation rate-matched unobserved variants as pathogenic, and utilizing 3D structural protein language models. A notable difference is that AlphaMissense uses predominantly human missense variants for training while most of PrimateAI-3D’s training data comes from NHP sequencing. In all real-world cohorts, PrimateAI-3D outperforms AlphaMissense including on both biobank benchmarks, which assess variant impacts on clinical phenotypes and blood protein levels, and on the three rare disease cohorts, which evaluate the classifier’s ability to distinguish mutations observed in patients with the diseases compared to a healthy control population. Among 31 genes tested with in vitro saturation mutagenesis assays, AlphaMissense performed better in 16 genes, while PrimateAI-3D performed better in 15 of the genes. AlphaMissense also performed slightly better on the ClinVar benchmark.

Overall, the good performance of AI models such as both PrimateAI-3D and AlphaMissense on multiple benchmarks suggests that deep-learning strategies are able to identify clinically and biologically relevant effects of missense variation and are a leading direction for future research in genetic variants study.

Multi-Omics Methodology

Besides using AI and machine learning as tools to process genomes data, multi-omics is another new sequencing methodology that combines various levels of genome complexity to integrate them and generate more insights. In the post-genomic era, a large list of “omics” technologies are developed, including epigenomics, transcriptomics, proteomics, and metabolomics, at both bulk and single-cell levels, each focusing on one distinct biology aspect of the genome. Each “omics” standing alone provides a limited view of human biology and diseases. This approach provides a more holistic view of health and disease, and a more comprehensive picture of gene functions and regulation in disease biology¹². Multi-



omics methods are also helping advance the development of precision medicine in a range of human diseases areas, such as cancer, neurodegenerative diseases, cardiovascular disease, immune health and women’s health¹³.

In addition to human genomes, NHP genome sequencing is also an essential dataset used in multi-omics methodology, where analyzing and examining how genetic variations affect gene expression and protein function in different NHP species are done from the multi omics lenses. For example, the prefrontal cortex (PFC) is known to be a key brain region responsible for age-related cognitive decline. However, little is known about aging-related molecular changes in PFC that may mediate these effects. One research group integrated multiple omics data types (transcriptomics, proteomics, metabolomics) from samples across the adult age span, in order to quantify PFC changes associated with healthy aging in female baboons. Their integrated omics approach used unbiased weighted gene co-expression network analysis to integrate data and treat age as a continuous variable, and revealed highly interconnected known and novel pathways associated with PFC aging. The research found Gamma-aminobutyric acid (GABA) tissue content associated with these signaling pathways, providing potential biomarkers to assess PFC changes with age. These highly coordinated pathway changes during aging may represent early steps for aging-related decline in PFC functions, such as learning and memory, and provide potential biomarkers to assess cognitive status in humans¹³.

Xenotransplantation

Organ, tissue and cell transplantation have been life-saving for end-of-life stage human patients, and due to the shortage of human organs for transplant, researchers have begun to turn to xenotransplant as an alternative. Xenotransplantation is the process of transplanting organs or tissues from one species to another¹⁴. Two key obstacles have prevented successful xenotransplantation: potential risk of the transmission of viruses between species, known as zoonotic infectious diseases; and

immune-mediated incompatibilities between species leading to organ rejection¹⁵.

NHP genome data is helpful in xenotransplantation research as well. One major finding is that, NHPs are not considered a good source donor for xenotransplantation to humans, despite their genetic close relations to humans, On the contrary, the close genetic distance between NHP and humans actually poses the weakest species barrier between donor and recipient species, resulting in highly likely interspecies transmission of pathogens and infectious diseases.

Recent innovations in gene-editing technology enable genetically engineered pigs to become more feasible organ donors in xenotransplantation¹⁶. Because of their phylogenetic distance from humans, the likelihood of cross-species transmission of infections with pig donors is less than between NHP and humans. Pig donors are often genetically engineered based on NHP genetic models. Also, before proceeding to human studies, a clinically ready porcine donor must be engineered and its xenograft successfully tested in NHP recipients. The study of NHP genomes sequencing data is instrumental in xenotransplantation research by providing insights into organ compatibility, disease transmission risks and immunological barriers.

Discussion

Restatement of Key Findings

The research has the following key findings and conclusions: First, accurate clinical interpretation of human genetic variants are crucial in effective precision medicine, helping predicting, preventing and treating diseases, as well as helping developing and designing optimized clinical trials for target drugs, and treating patients with individualized target drugs and dosage. The second finding directly answers the main question of this research, i.e. NHP genome data is indeed an essential valuable complementary data source combined with human genomes data to help analyze and identify pathogenic human genetic

variants. The main reasons for NHP genomes data being so valuable include NHP's genetic closeness to humans, as well as their having been through much longer natural selection, hence the common genetic variants found in both NHP and humans are most likely benign, these annotation can help classify and compare genetic variants, especially when used via advanced statistics technologies and machine learning AI technologies to process the data effectively.

The third finding reviews a concrete NHP genomes processing project PrimateAI-3D, which uses AI and statistical technologies to process a combination of NHP genomes data and multiple human genomes databases. The project successfully identified multiple pathogenic genetic variants in human diseases such as heart disease and mental illness diseases, with benchmark performance better than other similar AI projects not using NHP data. This example proves NHP data is valuable in helping understanding human diseases.

The research also includes brief overviews of NHP genome usages in multi-omics methodology and xenotransplantation research. NHP genome data analyzed using multi-omics methodology helps create more insight in molecular biology and human diseases. In the xenotransplantation process, NHP genomes study helps confirm NHP is not good to be used as xenotransplant sources due to its close genetic distance to humans, also NHP genetic model is used to design genetic engineering of pigs as donor sources.

Implications and Significance: The research helps confirm NHP genomes data is a valuable datasource to help identify pathogenic human genetic variants for academic research. Therefore, it is worthwhile to scale more comprehensive sequencing of NHP species, as well as further develop and optimize machine learning models and AI algorithms to process NHP genomes data effectively.

Connection to Objectives: The research started with the hypotheses that NHP genomes data is useful to shed light on further insights on how human genetic variants correlate with diseases. The hypothesis is met with convincing theory on its effective and practical successful AI project results.

Recommendations: Based on the research, further sequencing of more individuals in most or all NHP species are valuable and worthwhile for the science community to spend the resources and effort. It not only helps understand primates evolution history, but more comprehensive scaled NHP genomes data help increase clinical understanding of human genetic variants, as shown in the PrimateAI-3D project results.

Limitations: The limitation of the research is that only science documentation on relevant topics available from recent 2 years are reviewed. Also, other than PrimateAI-3D, most

other genomes research done via AI are trained with human genome data only, hence further peer review and similar AI projects processing NHP genome data would add more proof to the effectiveness of the data usage.

Closing Thought: NHPs are cousins of human beings in natural evolution history and many NHP species are in danger of extinction. Protecting and preserving NHP species, and sequencing of comprehensive NHP genomes, and more effective processing of NHP genomes data using AI technologies will help gain more insights on genetic history of primates including humans, and particularly improve clinical understanding of human genetic variants and help predict disease risks and candidate drug targets.

Methods

Search Strategy

Search has been done in respectable science journals such as Science, Nature, NEJM and official sources such as NIH and government health agency sites. Keywords searched include: genomes, human genetic variants, precision medicine, human genomes project, primates genomes, AI, machine learning on genomes data.

Inclusion Criteria

Science articles mostly from recent 2 years relevant on the topic are reviewed and included to ensure the information is up to date. Most documents are published peer reviewed in respectable international journals, some references are from NIH, NCBI or medRxiv which are sufficiently credible.

Data Extraction

Hypothesis and practical examples are consolidated and narrative is synthesized in the paper to explicitly address the original research questions in a logical order. Specific algorithm, research results data and architecture figures are directly cited from the published papers by the PrimateAI-3D project in Science journal.

Synthesis Method

The research is done after reviewing multiple research documents. Both thematic analysis and narrative synthesis methods are used to elucidate how both the similarity and differences between NHP genomes and human genomes help make the combination of both data as valuable source to identify clinical importance of human genetic variants, and how AI technology is used to help processing the enormous data effectively. The research consolidates these findings in a

logical order: first the context and problem—why it is valuable yet challenging to identify pathogenic human genetic variants; then the answer of adding NHP genomes data as valuable complementary datasource to help assess genetic variants' impact on diseases.

Quality Assessment

Only documents in credible international journals and official government science agency websites are used in the research to ensure the source is valid and reliable.

Acknowledgements

Major thanks to my teachers at Green Level High School who sparked my interest in science and helped me grow confidence in research and writing. Thanks to my mom who not only encouraged me to pursue my intellectual curiosity with reading and quests, but also helped review and provide improvement advice multiple times to this article. Thanks also to my dad for all the encouragement and support through the weeks I'm refining and editing the paper to ensure I aimed at a level of excellence at each step.

Author and affiliations

This research article is authored by Ethan Wan, a 11th grader in Green Level High School in Cary, NC. Ethan is enthusiastic in science and engineering fields and particularly in biochemistry, anthropology, mathematical and medical sciences. Ethan was interested in the human genome then primate's genome projects then began to read and research in this field and completed this article with his understanding and conclusions on how study of primate genomes can positively contribute to battle human disease and improve human health care.

References

- 1 A. B. J. A. L. J *et al.*, *Molecular Biology of the Cell*, Garland Science, New York, 4th edn, 2002.
- 2 National Human Genome Research Institute, *The Human Genome Project*, <https://www.genome.gov/human-genome-project>, 2020, Accessed: 22 Dec. 2020.
- 3 P.-M. Lledo, *Personalised, preventive, predictive, participatory: the 4Ps of tomorrow's medicine*, <https://www.polytechnique-insights.com/en/columns/health-and-biotech/personalized-preventive-predictive-participatory-the-4ps-of-tomorrows-medicine/>, 2024, Accessed: June 2024.
- 4 P. P. Fiziev *et al.*, *Science*, 2023.
- 5 *What is ClinVar*, <https://www.ncbi.nlm.nih.gov/clinvar/intro/>, 2024, Newest release accessed in 2024.
- 6 *What Do Primates Have in Common? Humans & Our Cousins*, <https://www.amnh.org/exhibitions/permanent/human-origins/understanding-our-past/living-primates>, 2020, Accessed: 2020, American Museum of Natural History.
- 7 H. Gao *et al.*, *Science*, 2023.
- 8 *How Does a Convolutional Neural Network Work?*, <https://www.nvidia.com/en-us/glossary/convolutional-neural-network/>, 2024, Accessed: 2024.
- 9 P. Fiziev, J. McRae, T. Hamp, H. Gao and K. Farh, *Improving genetic risk prediction and drug target discovery using primate DNA and advanced artificial intelligence*, <https://www.illumina.com/science/genomics-research/articles/primataei-3d.html>, 2023, Accessed: June 1, 2023, Illumina website.
- 10 D. A. Parry *et al.*, *medRxiv*, 2024.
- 11 K. Minton, *Nature*, 2023.
- 12 *Multi-Omics for Health and Disease (Multi-Omics)*, <https://www.genome.gov/research-funding/Funded-Programs-Projects/Multi-Omics-for-Health-and-Disease>, 2024, Accessed: 2024.
- 13 L. A. Cox *et al.*, *Neurobiology of Aging*, 2023, **132**, 109–119.
- 14 D. K. C. Cooper, *Baylor University Medical Center Proceedings*, 2012, **25**, 49–57.
- 15 M. D. Dooldeniya and A. N. Warrens, *JRSM*, 2003, **96**, 111–117.
- 16 R. P. Anand *et al.*, *Nature*, 2023, **622**, 393–401.